Artificial Intelligence, Values, and Alignment

Year: 2020 | Citations: 809 | Authors: Iason Gabriel

Abstract

This paper looks at philosophical questions that arise in the context of AI alignment. It defends three propositions. First, normative and technical aspects of the AI alignment problem are interrelated, creating space for productive engagement between people working in both domains. Second, it is important to be clear about the goal of alignment. There are significant differences between AI that aligns with instructions, intentions, revealed preferences, ideal preferences, interests and values. A principle-based approach to AI alignment, which combines these elements in a systematic way, has considerable advantages in this context. Third, the central challenge for theorists is not to identify 'true' moral principles for AI; rather, it is to identify fair principles for alignment that receive reflective endorsement despite widespread variation in people's moral beliefs. The final part of the paper explores three ways in which fair principles for AI alignment could potentially be identified.